

Solution 1: Risk Minimizers for Generalized L2-Loss

- (a) For the optimal constant model $f(\mathbf{x}) = \theta$ for the loss $L(y, f(\mathbf{x})) = (m(y) - m(f(\mathbf{x})))^2$, we first apply the following substitution $z^{(i)} = m(y^{(i)})$ for each $i = 1, \dots, n$, and introduce $\theta_m = m(\theta) \in m(\mathbb{R})$. Note that the inverse of m is continuous and strictly monotone as well, so that the minimizer of the initial optimization problem, i.e.,

$$\arg \min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg \min_{\theta \in \mathbb{R}} \sum_{i=1}^n (m(y^{(i)}) - m(\theta))^2.$$

is the same as for the “substituted” optimization problem, i.e.,

$$m^{-1} \left(\arg \min_{\theta_m \in m(\mathbb{R})} \sum_{i=1}^n (z^{(i)} - \theta_m)^2 \right).$$

For the term in the brackets we have seen in the lecture (optimizer of the empirical L2 risk) that

$$\arg \min_{\theta_m \in m(\mathbb{R})} \sum_{i=1}^n (z^{(i)} - \theta_m)^2 = \frac{1}{n} \sum_{i=1}^n z^{(i)} = \frac{1}{n} \sum_{i=1}^n m(y^{(i)}).$$

Consequently,

$$\hat{f}(\mathbf{x}) = m^{-1} \left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right)$$

is the optimal constant model for L .

- (b) First, note that

$$\begin{aligned} \mathcal{R}_L(\hat{f}) &= \mathbb{E}_{xy} [L(y, \hat{f}(\mathbf{x}))] \\ &= \mathbb{E}_{xy} [(m(y) - m(\hat{f}(\mathbf{x})))^2] \\ &= \mathbb{E}_{xy} \left[\left(m(y) - \frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right)^2 \right] \\ &= \mathbb{E}_{xy} [m(y)^2] - 2\mathbb{E}_{xy} \left[m(y) \frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right] + \mathbb{E}_{xy} \left[\left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \right]. \end{aligned}$$

Now, because $y, y^{(1)}, \dots, y^{(n)}$ are i.i.d. with $\mathbb{E}_{xy} [m(y^{(i)})] = \mathbb{E}_{xy} [m(y)]$, we get

$$\begin{aligned} \mathbb{E}_{xy} \left[m(y) \frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right] &= \frac{1}{n} \mathbb{E}_{xy} \left[m(y) \sum_{i=1}^n m(y^{(i)}) \right] \\ &= \frac{1}{n} \mathbb{E}_{xy} [m(y)] \mathbb{E}_{xy} \left[\sum_{i=1}^n m(y^{(i)}) \right] \\ &= \frac{1}{n} \mathbb{E}_{xy} [m(y)] n \mathbb{E}_{xy} [m(y)] = \mathbb{E}_{xy} [m(y)]^2. \end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}_{xy} \left[\left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \right] &= \frac{1}{n^2} \left(\sum_{i=1}^n \mathbb{E}_{xy} \left[m(y^{(i)}) \left(\sum_{i=1}^n m(y^{(i)}) \right) \right] \right) \\
&= \frac{1}{n^2} \left(\sum_{i=1}^n \mathbb{E}_{xy} \left[m(y^{(i)})^2 + \sum_{j \neq i} m(y^{(i)}) m(y^{(j)}) \right] \right) \\
&= \frac{1}{n^2} \left(\sum_{i=1}^n \mathbb{E}_{xy} [m(y^{(i)})^2] + \sum_{i=1}^n \sum_{j \neq i} \mathbb{E}_{xy} [m(y^{(i)}) m(y^{(j)})] \right) \\
&= \frac{1}{n^2} \left(n \mathbb{E}_{xy} [m(y)^2] + \sum_{i=1}^n \sum_{j \neq i} \mathbb{E}_{xy} [m(y^{(i)})] \mathbb{E}_{xy} [m(y^{(j)})] \right) \\
&= \frac{1}{n^2} \left(n \mathbb{E}_{xy} [m(y)^2] + n(n-1) \mathbb{E}_{xy} [m(y)]^2 \right) \\
&= \frac{1}{n} \mathbb{E}_{xy} [m(y)^2] + \left(1 - \frac{1}{n}\right) \mathbb{E}_{xy} [m(y)]^2.
\end{aligned}$$

So, combining the three later math displays, we obtain

$$\begin{aligned}
\mathcal{R}_L(\hat{f}) &= \mathbb{E}_{xy} [m(y)^2] - 2 \mathbb{E}_{xy} \left[m(y) \frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right] + \mathbb{E}_{xy} \left[\left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \left(\frac{1}{n} \sum_{i=1}^n m(y^{(i)}) \right) \right] \\
&= \mathbb{E}_{xy} [m(y)^2] - 2 \mathbb{E}_{xy} [m(y)]^2 + \frac{1}{n} \mathbb{E}_{xy} [m(y)^2] + \left(1 - \frac{1}{n}\right) \mathbb{E}_{xy} [m(y)]^2 \\
&= \left(1 + \frac{1}{n}\right) \left(\mathbb{E}_{xy} [m(y)^2] - \mathbb{E}_{xy} [m(y)]^2 \right) \\
&= \left(1 + \frac{1}{n}\right) \text{Var}(m(y)).
\end{aligned}$$

- (c) In order to derive the risk minimizer, we consider the unrestricted hypothesis space $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathbb{R}\}$. By the law of total expectation

$$\begin{aligned}
\mathcal{R}_L(f) &= \mathbb{E}_{xy} [L(y, f(\mathbf{x}))] \\
&= \mathbb{E}_x [\mathbb{E}_{y|x} [L(y, f(\mathbf{x})) \mid \mathbf{x}]] \\
&= \mathbb{E}_x [\mathbb{E}_{y|x} [(m(y) - m(f(\mathbf{x})))^2 \mid \mathbf{x}]].
\end{aligned}$$

Since \mathcal{H} is unrestricted we can choose f as we wish: At any point $\mathbf{x} = \mathbf{x}$ we can predict any value c we want. The best point-wise prediction is

$$f^*(\mathbf{x}) = \underset{c}{\text{argmin}} \mathbb{E}_{y|x} [(m(y) - m(c))^2 \mid \mathbf{x}] \stackrel{(*)}{=} m^{-1}(\mathbb{E}_{y|x} [m(y) \mid \mathbf{x}]),$$

where $(*)$ is due to

$$\begin{aligned}
\underset{c}{\text{argmin}} \mathbb{E} [(m(y) - m(c))^2] &= \underset{c}{\text{argmin}} \underbrace{\mathbb{E} [(m(y) - m(c))^2] - (\mathbb{E}[m(y)] - m(c))^2}_{=\text{Var}[m(y) - m(c)] = \text{Var}[m(y)]} + (\mathbb{E}[m(y)] - m(c))^2 \\
&= \underset{c}{\text{argmin}} \text{Var}[m(y)] + (\mathbb{E}[m(y)] - m(c))^2 = m^{-1}(\mathbb{E}[m(y)]),
\end{aligned}$$

because $\text{Var}[m(y)]$ does not depend on c . Note that we could have used a similar substitution as in (a) here to derive f^* . Furthermore, if we use $m(x) = x$ such that the considered loss coincides with the L2 loss, we get (quite naturally) the same best point-wise prediction as for the L2 loss. Using an m corresponding to another notion of mean (e.g., harmonic or geometric mean), the best point-wise prediction for that other mean is obtained in each case.

- (d) The optimal constant model in terms of the (theoretical) risk can be obtained from the previous by forgetting the conditioning on point $\mathbf{x} = \mathbf{x}$, which leads to

$$\bar{f}(\mathbf{x}) = m^{-1}(\mathbb{E}_y [m(y)]).$$

The risk of the latter is $\text{Var}(m(y))$:

$$\mathcal{R}_L(\bar{f}) = \mathbb{E}_{xy}[(m(y) - m(\bar{f}(\mathbf{x})))^2] = \mathbb{E}_y[(m(y) - \mathbb{E}_y[m(y)])^2] = \text{Var}(m(y)).$$

(e) The Bayes regret can be decomposed as follows:

$$\mathcal{R}_L(\hat{f}) - \mathcal{R}_L^* = \underbrace{\left[\mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f) \right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^* \right]}_{\text{approximation error}}.$$

If we consider as the hypothesis space $\mathcal{H} = \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f(\mathbf{x}) = \theta \ \forall \mathbf{x} \in \mathcal{X}\}$, i.e., the set of constant models, then the estimation error is

$$\begin{aligned} \mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f) &= \underbrace{\mathcal{R}_L(\hat{f})}_{\stackrel{(b)}{=} (1 + \frac{1}{n}) \text{Var}(m(y))} - \underbrace{\mathcal{R}_L(\bar{f})}_{\stackrel{(d)}{=} \text{Var}(m(y))} = \left(1 + \frac{1}{n}\right) \text{Var}(m(y)) - \text{Var}(m(y)) = \frac{1}{n} \text{Var}(m(y)), \end{aligned}$$

while the approximation error is

$$\begin{aligned} \inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^* &= \underbrace{\mathcal{R}_L(\bar{f})}_{\stackrel{(d)}{=} \text{Var}(m(y))} - \mathcal{R}_L(f^*) \\ &= \text{Var}(m(y)) - \mathbb{E}_x \left[\mathbb{E}_{y|x} [(m(y) - m(f^*(\mathbf{x})))^2 \mid \mathbf{x}] \right] \\ &= \text{Var}(m(y)) - \mathbb{E}_x \left[\mathbb{E}_{y|x} \left[(m(y) - m(m^{-1}(\mathbb{E}_{y|x}[m(y) \mid \mathbf{x}]))^2 \mid \mathbf{x} \right] \right] \\ &= \text{Var}(m(y)) - \mathbb{E}_x \left[\underbrace{\mathbb{E}_{y|x} [(m(y) - \mathbb{E}_{y|x}[m(y) \mid \mathbf{x}])^2 \mid \mathbf{x}]}_{= \text{Var}[m(y) \mid \mathbf{x}]} \right] \\ &= \text{Var}(m(y)) - \mathbb{E}_x [\text{Var}[m(y) \mid \mathbf{x}]] \\ &= \text{Var}(\mathbb{E}_{y|x}[m(y) \mid \mathbf{x}]). \end{aligned}$$

Note that the larger the sample size n the lower the estimation error, while the approximation error remains constant.